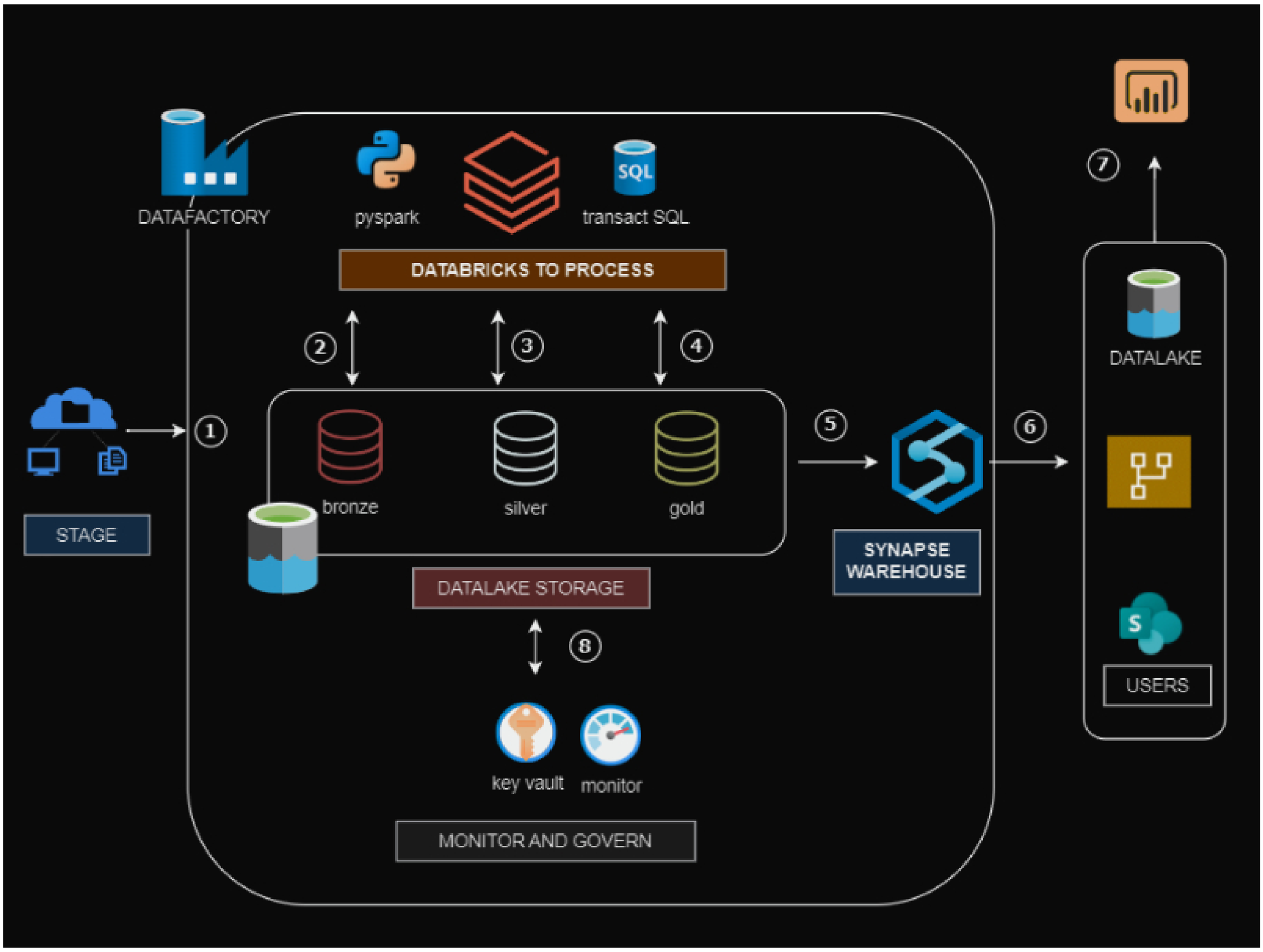
Report system Documentation

Creator **technology @goglt.com** Created **Jun 27, 2024, 09:14** Last updated **Sep 13, 2024, 15:57**



lakehouse view

Tools

Datafactory

In this project, we use Azure Data Factory to orchestrate the data workflows, ensuring that all Databricks notebooks are executed for necessary transformations and all Synapse stored procedures are run to load tables. Two distinct pipelines in Data Factory cater to different refresh frequencies: one pipeline refreshes tables six times a day, while the other handles tables that require a daily refresh. This setup optimizes data processing and ensures timely updates across the system. In the lakehouse view, Data Factory executes processes 1 through 5.

Databricks

This tool manages the entire data processing and transformation workflow, enabling the implementation of the medallion architecture within Databricks. The medallion architecture is a data design pattern used to logically organize data in a lakehouse, with the goal of incrementally enhancing the structure and quality of the data as it progresses through each layer—Bronze, Silver, and Gold.

- In the **Bronze layer**, data is ingested in its raw form, mirroring the source. This layer typically contains all columns as strings, and data is often in formats like CSV.
- The **Silver layer** introduces transformations, such as selecting relevant columns, renaming them, and defining appropriate data types. This layer refines the data, making it more usable.
- Finally, the **Gold layer** contains data that has been fully transformed and is ready for consumption by end-users. This data is shared with various systems, including Power BI, file systems, and applications.

Synapse

This tool functions as a serverless data warehouse. In this environment, data is stored in structured tables, ensuring efficient and secure connectivity through Transact-SQL. Azure Synapse provides the capability to create stored procedures, enabling dynamic updates to all tables and maintaining a responsive warehouse. This serverless warehouse leverages data lakes for storage and uses code to recreate and manage tables as needed.

In this serverless data warehouse setup, data lakes are used as the primary storage layer. Data lakes are scalable and cost-effective storage solutions that can handle large volumes of raw data in various formats. Instead of storing data in a traditional database, the warehouse leverages the flexibility of data lakes.

To manage and use this data effectively, code (typically SQL or Spark) is employed to dynamically recreate and update the tables within the warehouse. This approach allows for on-demand table creation and transformation directly from the data stored in the lake, enabling a flexible and scalable data architecture that can adapt to changing data needs without the constraints of a fixed schema

Data Lakes

Data lakes serve as our primary storage resource, holding all data in various formats. We organize the data into Bronze, Silver, and Gold containers to facilitate transformations using Databricks. These containers align with the medallion architecture, allowing us to progressively refine the data. The same containers are then utilized in Azure Synapse to dynamically recreate tables, ensuring consistency and seamless integration across our data processing workflows.

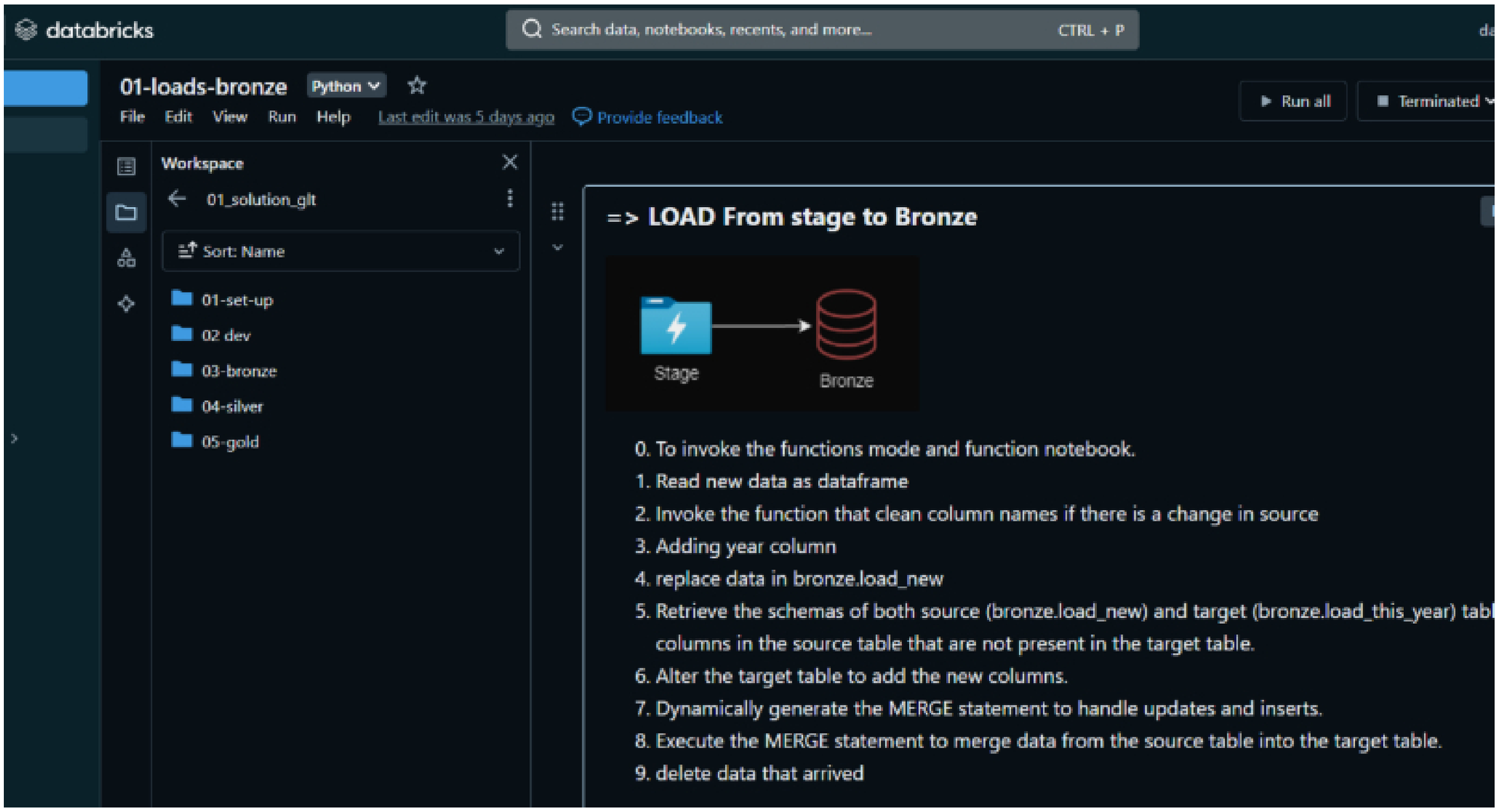
Workflow

1 Stage

Various data ingestion methods bring data into a staging area before it reaches the Bronze containers. Tools like Dataflow, Power Automate, Google Functions, Azure Functions, and others are used to capture and load data into this pre-Bronze stage. Once the data is in the staging area, it is then moved into the Bronze containers for further processing and transformation

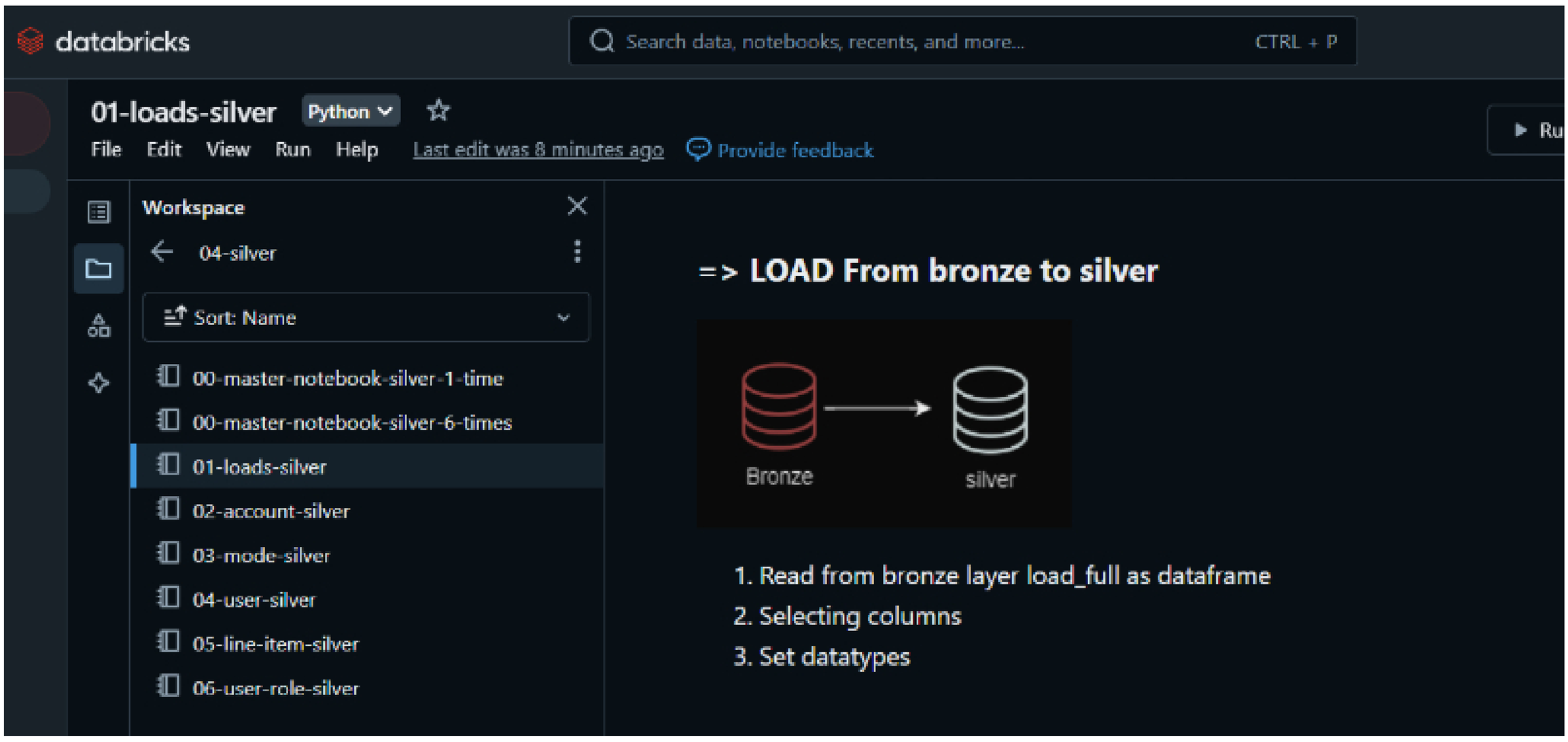
2 Bronze layer

In the Bronze layer, data is ingested in its raw form, directly reflecting the source. This layer typically stores all columns as strings, with data often in formats like CSV. A powerful function in Databricks, called UPSERT, is used here, particularly with Delta tables. This function allows for the efficient creation of smaller, incremental tables to bring new data into a target table. The lastModifiedDate column serves as the key, enabling the update of rows that match between the source and target tables, while also inserting new data when no match is found. The data is stored in Azure data lake container called Bronze, where we keep all the raw, original data



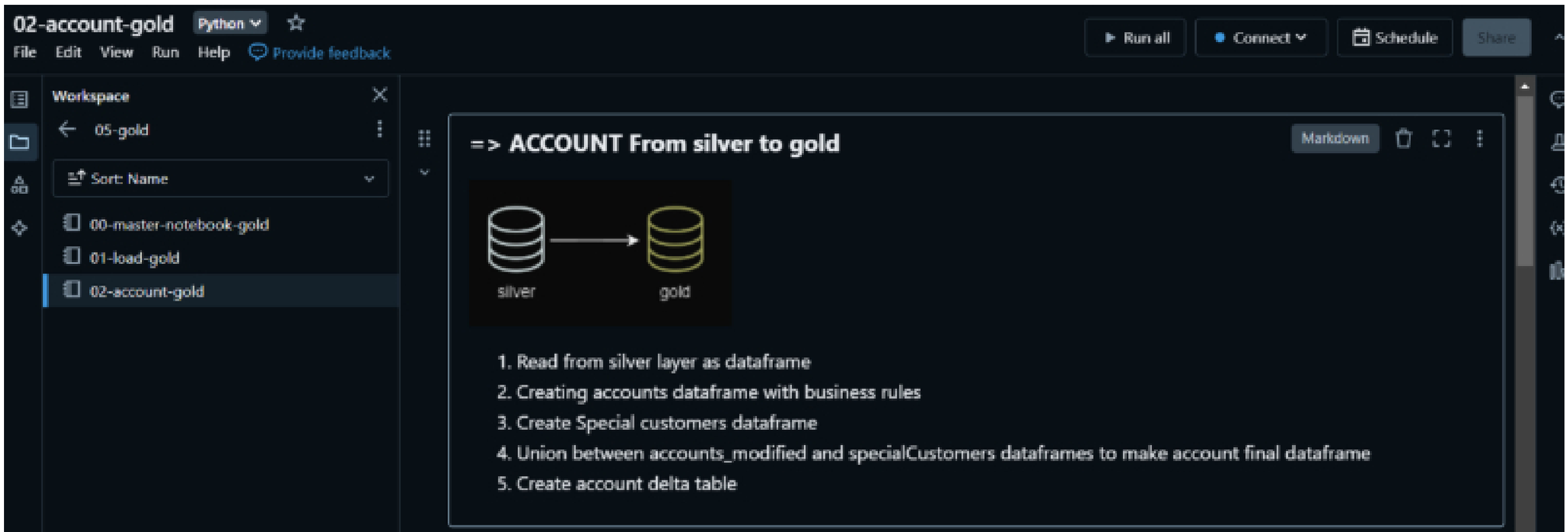
3 Silver layer

The data comes from the bronze layer, and we use Databricks PySpark to do all the basic transformation: changing data types, dropping unnecessary columns, and renaming columns, etc Then, the data gets stored in Azure data lake container called Silver, where we keep it after these initial transformations.



4Gold layer

The data come from the Silver layer, and we use Databricks PySpark for more advanced transformations:
Joins, group by, unions, control structures (like if statements), arithmetic, and more. After that, the data is stored in Azure data lake container called Gold, ready for analytics tools like Power BI to use.

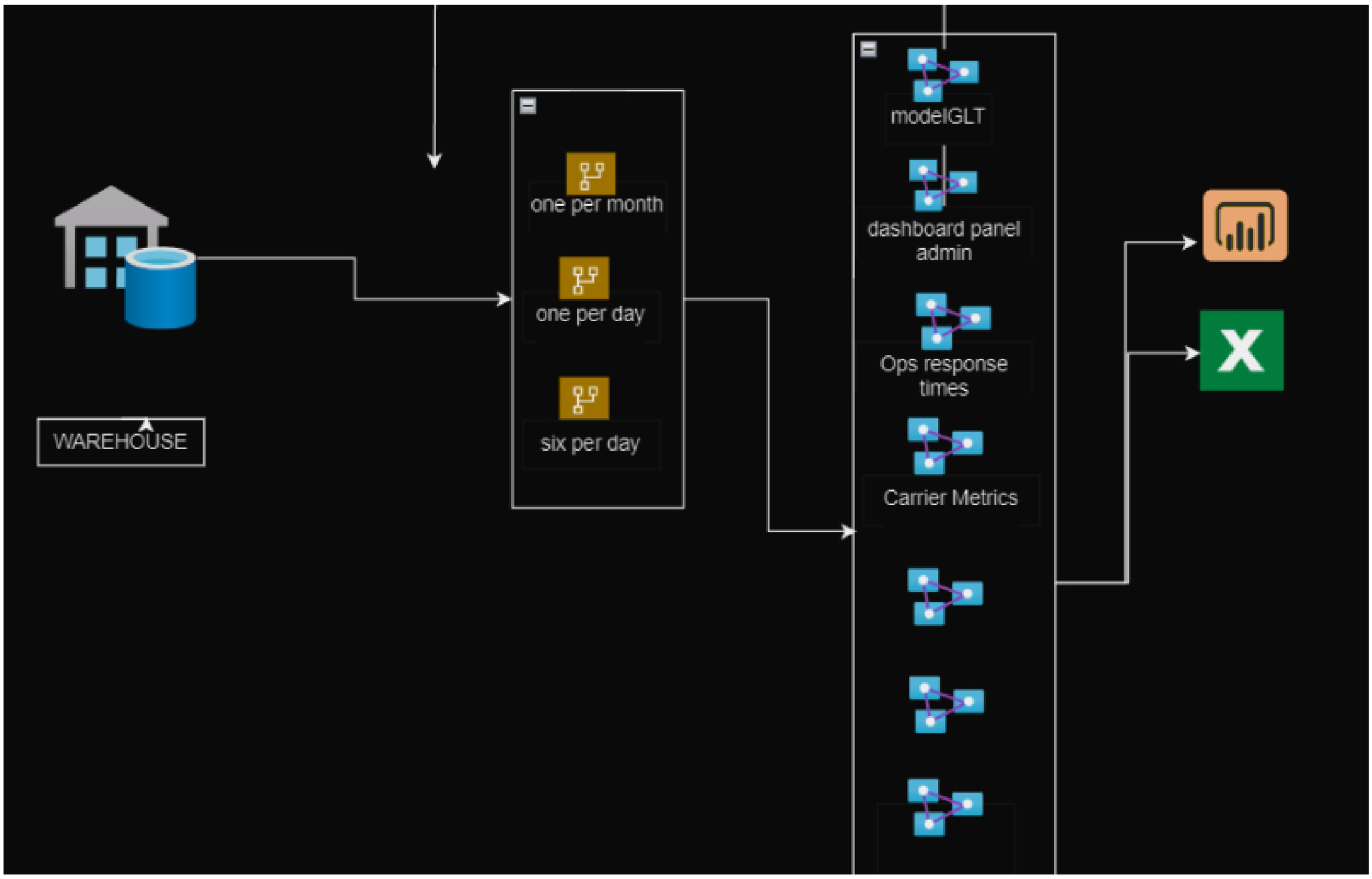


5Warehouse

With Azure Synapse Analytics, we can build a fully serverless data warehouse by leveraging the Bronze, Silver, and Gold containers from our data lake. Using SQL scripts, we create external tables that directly access the data in these containers without needing to move it. The Silver layer holds cleaned and transformed data, while the Gold layer contains fully refined, ready-to-use data. This setup allows for efficient querying and analysis within the warehouse, all while keeping the data in the lake, making it ideal for seamless integration with analytics tools such as Power Bi.

6Virtual data with dataflows , data lakes or sharepoints

In order to reduce costs from the serverless warehouse, which charges on demand, the data is virtualized using dataflows for BI environments, SharePoint for apps like Excel, and data lakes for web development. This way, we can create as many models as we need without increasing costs, as well as spreadsheets and apps. This data virtualization across these three storage points acts as a controllers between Synapse and analytics tools like Power BI.



7Reporting

With clean data, we perform advanced analysis using DAX and create visualizations in Power BI.

8Monitor and Govern

Azure Key Vault is used to protect SAS tokens and credentials needed to access data stored in containers. Meanwhile, Azure Monitor, a native tool in the Azure portal, helps track daily billing.